Prairie Dogs
Supplement 3.40

Statistics for Data Analysis and Hypothesis Testing

Notes

Mathematics is the language of science. All sciences require some degree of proficiency in mathematics. A branch of mathematics used extensively in **all** sciences is **statistics.**

Many kinds of biological observations consist of numerical information called data. Statistics provide an objective method for one to use to present and analyze research data.

For the purposes of this unit, only the terms and statistics needed to analyze data from the project will be introduced.

Definitions of Key Terms

Sample Any small group of individuals or objects selected to represent the entire group called the population. The letter **n** is used to represent the sample size number.

Population Any set of individuals or objects having some common observable characteristics. In biology, by definition, species form populations. There are two kinds of populations -- finite and infinite. Most of the populations in biology are infinite.

Parameter A numerical measurement describing some characteristic of a population.

Statistics A numerical measurement describing some characteristic of a sample.

Descriptive Statistics Numerical, graphical and tabular methods for organizing and summarizing data.

Inferential Statistics Methods for generalizing from a sample to draw conclusions and make decisions or predictions about a population. To be valid, samples have to be obtained randomly (without bias).

Random Sample (of size **n**) Sample selected in such a way that gives every different sample of **n** an equal chance of being selected. (**n** represents the total number of data sampled for a given population.)

Measures of Central Tendency

Most of us are familiar with some of the measures. The most widely used measure of central tendency is the **arithmetic mean** or average value of a set of numerical data.

Other measures are mode, median and midpoint. All of these measures are used to locate the center of a distribution of a set of data.

For our purposes, we are only going to consider the sample mean, which is the most efficient, unbiased and consistent estimate of the population mean.

The sample mean, denoted as **X** (called *x-bar*) of a set sample values $x_1 + x_2 + x_3, \ldots, x_n$, where **n** is the sample size, is given by

$$\bar{X} = \frac{x_i}{n}$$

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\text{(sum of the data)}}{\text{(number of pieces of data)}}$$

WORK EXAMPLE PROBLEM.

From the example problem, notice that the sample contains identical data for several values of the variable. When this occurs, it may be convenient to record the data in the form of a frequency table. A frequency distribution is a way of grouping data so meaningful patterns can be found. **Frequency** means *how often* the same numerical value occurs. Now, let x_i denote each different measurement and f_i denote the frequency that each x_i occurs in the sample. The sample mean can now be calculated by

$$\bar{X} = \frac{x_i f_i}{n} = \frac{\text{(sum of the products)}}{n}$$

WORK EXAMPLE PROBLEM USING FREQUENCY TABLE.
PLOT THE DATA AS A HISTOGRAM.

To make a **histogram** (a kind of bar graph) plot x_i vs. f_i . The mean is located at the center of gravity of the histogram.

STUDENTS DO MEASURES OF CENTRAL TENDENCY WORK SHEET.

Measures of Dispersion

Measure of dispersion (also called measure of variability) is an indication of the scatter of the measurements around the center of the distribution. It also can indicate how clustered the measurements are around the center of the distribution. Some measures of dispersion are range, mean deviation, variance, standard deviation and standard error of the mean, which is also referred to as standard error.

Since variance and standard deviation are mathematically related, an understanding of these two statistics is necessary before discussing the standard error of the mean**.** These measures are based on a normal distribution (bell-curve) and for a population that is not normal, given a large enough sample size, the distribution approaches a normal distribution.

The **sample variance** (s^2)is a measure of the spread of numbers (values) about the sample mean and is given by

$$s^2 = \frac{(x_i - \bar{X})^2}{n - 1} = \frac{x_i^2 - \frac{(x_i)^2}{n}}{n - 1}$$

The sample standard deviation is the positive square root of the sample variance and is given by

$$s = \sqrt{s^2} = \sqrt{\frac{(x_i - \bar{X})^2}{n - 1}} = \sqrt{\frac{x_i^2 - \frac{(x_i)^2}{n}}{n - 1}}$$

The standard deviation of the sampling distribution of the mean is referred to as the **standard error of the mean** (s_x) and is given by

$$s_x = \frac{s}{\sqrt{n}}$$

If random samples of size **n** are from a normal population, the means of these samples will form a normal distribution (bell-curve). The distribution of means from a non-normal population will approach normality as the sample size **n** increases. This means that if you sample a population that is not normal, by increasing the sample size you can, in essence, have a normal distribution according to the Central Limit Theorem.

WORK EXAMPLE PROBLEMS.
STUDENTS DO MEASURES OF DISPERSION WORK SHEET.

Hypothesis Testing and the Chi-square Distribution

Hypothesis testing is an important branch of inferential statistics. In simple terms, a hypothesis is an educated guess or assumption about one or more of the population parameters that will either be accepted or rejected on the basis of the information obtained from a sample. There are several statistical methods available for hypothesis testing, but we are only going to look at the chi-square statistic.

A plant geneticist grows 100 progeny from a cross that is hypothesized to result in a 3 :1 phenotypic ratio of yellow-flowered to green-flowered plants. The cross produces a ratio of 84 yellow to 16 green-flowered plants. Based on the hypothesis, one would expect or predict a ratio of 75 yellow to 25 green-flowered plants. Now one has to determine whether the observed frequencies **deviate significantly** from the frequencies expected if the hypothesis were true.

The procedure for analysis of this type of problem begins with a concise statement of the hypothesis to be tested. As you recall, the hypothesis stated that the population of yellow-flowered to green-flowered plants was 3 : 1. In statistics this is referred to as a **null hypothesis** ($\mathbf{H_0}$). A null hypothesis is a statement of "no difference"; in this instance it means that the flower color population is not different from a 3 : 1 ratio. If $\mathbf{H_0}$ is found to be false, then an **alternate hypothesis** ($\mathbf{H_A}$) is assumed to be true. In this example, $\mathbf{H_A}$ would be that the flower population sampled has a flower color ratio that is not 3 yellow : 1 green. One always states a null hypothesis and an alternate hypothesis for every statistical test performed. All possible outcomes are accounted for by the two hypotheses.

The **chi-square** ($\mathbf{X^2}$) statistic can be used to measure how far a sample distribution deviates from a theoretical distribution:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The sample distribution frequency is represented by $\mathbf{O_i}$ (observed values in class **i**), and the theoretical distribution frequency is $\mathbf{E_i}$ (expected value of class **i**) if the null hypothesis is true, and is summed over all **k** categories data. In this sample, there are two categories of data (**k** = 2): yellow and green-flowered plants. The expected frequency $\mathbf{E_i}$ of each class is calculated by multiplying the total number of observations, **n**, by the proportion of the total that the null hypothesis predicts for the class. So, $\mathbf{E_1}$ = 100 x ¾ = 75 and $\mathbf{E_2}$ = 100 x ¼ = 25.

From the $\mathbf{X^2}$ equation it is apparent that larger differences between observed and expected frequencies will result in a larger $\mathbf{X^2}$ value, and smaller differences will result in a smaller $\mathbf{X^2}$ value. This type of calculation is referred to as a measure of **goodness of fit.** Goodness of fit means how close do my observed frequencies come to the expected frequency distribution? An $\mathbf{X^2}$ value of zero would indicate a perfect fit, therefore the smaller the $\mathbf{X^2}$ value is, the better the fit. Obviously, $\mathbf{X^2}$ can never be negative!

The words **deviate significantly** were used earlier in conjunction with hypothesis testing. From the $\mathbf{X_2}$ value as a measure of disagreement between observed and expected frequencies, a numerical value based on probabilities can be found to indicate the **statistical significance** of the disagreement. If the null hypothesis is true, is it likely (probable) to obtain an 84 : 16 flower color ratio from a random sample of this

population. If such a sample ratio can occur reasonably often, then there is no reason to reject $\mathbf{H_o}$.

But, if there is little chance (probability) of having an 84 : 16 flower color ratio from a random sample from a population having a 3 : 1 ratio, then one can conclude that the $\mathbf{H_O}$ is false and that the $\mathbf{H_A}$ is true. In biological work it is common to conclude that a $\mathbf{X^2}$ value with an associated probability of 5% or less is an indication that the $\mathbf{H_O}$ is false. The probability used as the criterion for rejection of the $\mathbf{H_O}$ is called the **significance level** (α **- alpha**), and α is selected before analyzing the data to eliminate bias (α **=** 5% in this example, but an α **=** 1% is sometimes used). The $\mathbf{X^2}$ value associated with the significance level is called the **critical value** of the statistic.

In chi-square, **degrees of freedom** (**DF = v = k -1)**, means that, given the frequencies in any **k -1** of the categories, one can calculate the frequency in the remaining category. This is true because **n** is known, and the sum of the frequencies in all **k** categories equals **n** (one has the freedom in assigning frequencies to only **k -1** of the categories).

Now back to the example problem.

OVERHEAD OF PROBLEM AND $\mathbf{X^2}$ TABLE OF CRITICAL VALUES.

Calculation of chi-square gives a value of 4.320. Now use the $\mathbf{X^2}$ table of critical values to find this value along the row where **v** = 1 (DF). This value lies between α = 0.05 and α = 0.025. So, for this problem one can state that $0.025 < P\ (\mathbf{X^2} \geq 4.302) < 0.05$., which is usually written as $0.025 < P < 0.05$. Since the $\mathbf{X^2}$ value is less than an α of 5%, the $\mathbf{H_O}$ is rejected and $\mathbf{H_A}$ is inferred to be true.

Prairie Dogs
Supplement 3.41

Measure of Central Tendency Example Problems

Transparency

A sample from a population of butterfly wing lengths.

X_t (cm)	X_t (cm)
3.3	4.0
3.5	4.0
3.6	4.0
3.6	4.1
3.7	4.1
3.8	4.1
3.8	4.2
3.8	4.2
3.9	4.3
3.9	4.3
3.9	4.4
4.0	4.5

$$\mathbf{X_t = 95.0cm}$$

$$\mathbf{n = 24}$$

$$\mathbf{\bar{X} = \frac{X_t}{n} = \frac{95.0\ cm}{24} = 3.96cm}$$

Prairie Dogs
Supplement 3.42

Measures of Central Tendency

Work Sheet

According to a study conducted by the US Geological Survey, concentrations of sulfate and nitrate, the two components of acid rain, declined significantly between 1980 and 1991. (Source: *Time*, July 19, 1993, p.20). One study of 20 lakes and rivers found an average concentration of 33, 42, 37, 53, 47, 41, 55, 38, 29, 38, 45, 53, 58, 27, 45, 52, 31, 46 and 32 units in each of these lakes or rivers.

1. Calculate the mean of the 20 sample means.

2. Draw the histogram for these sample means.

Prairie Dogs
Supplement 3.43

Measures of Central Tendency

Work Sheet Key

Solution to Problem #1

$$\text{Mean } \bar{X} = \frac{33 + 42 + 37 + \ldots = 31 + 46 + 32}{20} = \frac{840}{20} = 42$$

Solution to Problem #2

Prairie Dogs
Supplement 3.44

Measure of Dispersion Example Problems

Transparency

The calculation of the standard error of the mean, $\mathbf{s_x}$.

The following are data for systolic blood pressure, in **mm** of mercury.

121
125
128
134
136
138
139
141
144
145
149
151

$X = 1651mm$

$X^2 = 228{,}111mm^2$

$n = 12$

$$\mathbf{\bar{X} = \frac{1651mm}{12} = 137.6mm}$$

$$\mathbf{SS = 228{,}111mm^2 - \frac{(1651mm)^2}{12}}$$

$$\mathbf{= 960.9167mm^2}$$

$$\mathbf{s^2 = \frac{960.9167mm^2}{11} = 87.3561mm^2}$$

$$\mathbf{s = \sqrt{87.3561mm^2} = 9.35mm}$$

$$\mathbf{s_X = \frac{s}{\sqrt{n}} = \frac{9.35mm}{\sqrt{12}} = 2.7mm}$$

Prairie Dogs
Supplement 3.45

Measures of Dispersion
Work Sheet

1. State officials claim that the average amount of all charitable contributions during1999 was $964 with a standard deviation of $102. Many samples of size 64 are taken. Find the mean of these samples and the standard error of the mean.

2. Refer to previous exercise. What would the sample mean and standard error of the mean be if the samples are of (a) size 49 each, and (b) size 100 each.

Prairie Dogs
Supplement 3.46

Measures of Dispersion
Work Sheet Key

#1

$\bar{X} = \$964$

$s_x = \frac{102}{\sqrt{n}} = \frac{102}{\sqrt{64}} = \$12.75$

#2a

$\bar{X} = \$964$

and $s_x = \frac{102}{\sqrt{49}} = \$14.5714$

#2b

$\bar{X} = \$964$

and $s_x = \frac{102}{\sqrt{100}} = \$10.20$